ERIC Document Resume	See Instruction Sheet	DATE OF RESUME 6/3/65	DATE RECEIVED

A. PUBLICATION DATA

1. ERIC NUMBER	2. ORIGINATOR CONTROL NO(S).	3. IDENTIFICATION NO.
ED000666	8110	

DO NOT WRITE IN SPACE BELOW

4. AUTHOR(S): Harmon, Lindsey R.

5. TITLE OF DOCUMENT ABSTRACTED: The Significance of Descriptive Comments on Confidential Reports (Technical Report No. 7)

6. PUBLICATION TITLE: N.A.

7. EDITOR(S): N.A.

8. PLACE OF ORIGIN: ORG. OSP, NAS-NRC; ST. Washington, DC; CITY, STATE

9. PUBLISHER OR SOURCE: Office of Scientific Personnel, National Academy of Sciences - Nat. Research Council

10. VOL.	11. PAGES	12. MO./YR.	13. TOTAL PP.
N.A.	N.A.	Aug. 55	10

TELE. EXT. INST/LOC CODE

MICROFILM COPY AVAILABLE: YES ☒ NO ☐

COPYRIGHT GRANTED: YES ☐ NO ☒

COPYRIGHT RELEASE FILED: YES ☐ NO ☐

TYPE OF REPORT: PROPOSAL ☐ IN-PROGRESS ☐ FINAL ☒ FOLLOW-UP ☐

OTHER ERIC REFERENCES:

ERIC SATELLITE CODE

B. INDEXING DATA

14. SUBJECT AREA

15. TYPE OF ENDEAVOR: experimental study

16. INDEXING TERMS (6-12)

experimental studies, fellowship programs, measurement, higher education, financial aid, evaluation, graduate education, students

17. ABSTRACT (MAX. 280 WORDS)

A study to determine the reliability and eventually the validity of written confidential reports on students applying for National Science Foundation Fellowships. The three questions posed are: Can a descriptive report score be derived reliably and economically? Do confidential reporters agree among themselves? Do these scores yield unique information? Details of the process by which reliability is correlated are described in full. Conclusions indicated that a reliable and economic scoring system could be developed but individual reporters had a very low correlation with one another. It was felt that the Descriptive Report score, in all probability, does measure something unique and significant.

ERIC No.

18. SOURCE OF ABSTRACT: Reporter No. 6

C. FUNDING DATA

19. SUPPORTING ORGANIZATION	20. DIVISION OR DEPT.	21. CONTRACT NO. / PROJECT NO.
	22. TOTAL COST OF PROJECT	
	23. DURATION OF PROJECT	U.S. Department of Health, Education, and Welfare, Office of Education

ERIC Form 1

R-6-12
X 21

8110

TECHNICAL REPORT NO. 7
12 August 1955

Research on
Fellowship Selection Techniques
Supported by the
National Science Foundation

THE SIGNIFICANCE OF DESCRIPTIVE COMMENTS ON CONFIDENTIAL REPORTS

by Lindsey R. Harmon

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education

This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

Office of Scientific Personnel
National Academy of Sciences-
National Research Council
Washington, D. C.

ED000666

BRIEF

To insure that means used to select fellows for National Science Foundation fellowships are working to maximum effectiveness, a research program has been initiated to subject the selection techniques to the acid test of empirical validation. That is, each must be proved experimentally to pick out the people who will later be considered effective scientists. Because of the extensive time lag in determining relative degrees of on-the-job effectiveness, after entrance to graduate training, the earliest available applicants whose selection was made on techniques comparable to those now in use are now being followed up. These are the 1949 Atomic Energy Commission fellowship applicants, and both the awardees and rejectees are included in the experimental study.

One of the variables under study is a quantitative score derived from evaluation of the free-response descriptive comments on the Confidential Report Form. Three questions were asked with regard to this descriptive report score, and quantitative answers were found for each of them. The questions were:

a) Can such a score be derived reliably and economically?
b) Do Confidential Reporters agree among themselves in the evaluations which are indicated by these scores?
c) Do these scores yield unique information, i.e., something not already indicated by the simpler graphic rating scales?

The answer to the first question is that intelligent, college-educated clerical workers, given adequate training, can score these comments with sufficient reliability to make the measures meaningful and experimentally useful. However, separate scoring by three clerks is necessary to derive an average score of high reliability, and there could be no assurance that the necessary quality of scoring could be maintained in a large-scale operation such as the National Science Foundation fellowship program. The cost and calendar time delay in such scoring would also make it feasible only for research purposes.

Secondly, Confidential Reporters agree among themselves in the scorable information they put into descriptive comments as well as they do in the check marks they place on a graphic rating scale. The degree of agreement in both cases is rather low, however, as is usually found with such personal evaluations. The reliability of an average of the scores from three or four Confidential Reporters is, nevertheless, sufficiently high to justify their use as selection variables.

The answer to the third question is that the descriptive report scores probably yield unique information. The statistical evidence is not completely conclusive, but the probability is high enough to warrant carrying out further studies with this descriptive report score. It will be validated as a predictor of on-the-job effectiveness of young scientists, along with the graphic rating score and the scores derived from objective tests. Only thus can its ultimate usefulness be determined, and its proper weight assigned in the whole complex of evaluation information.

The Significance of Descriptive Comments On Confidential Reports

The improvement of techniques used in the selection of fellows for the National Science Foundation Fellowships is the general objective of a research program conducted by the National Academy of Sciences-National Research Council. Although there are many ways in which improvement may be made, the ultimate test of any technique is that of validity in selection of scientists. That is, each technique must be proved to be able to sort out those who will later be found to be most effective on the job as scientists from those who will be found to be not so effective. Such evaluation of selection instruments against a criterion of on-the-job effectiveness has been initiated. Because several years are required from the entrance to graduate school until graduation and establishment of one's effectiveness as a scientist, the time lag in such validation is large. Six years has been considered to be a minimum, and a good case can be made for the point that many more years, or even a whole lifetime, is necessary in order to establish such a criterion reliably and free from bias. As a practical matter, however, if progress is to be made in validation of selection instruments, a first approximation to such a criterion must be established in the first few years following graduation. In preparation for validation against such a criterion, a series of variables were assembled regarding applicants for Atomic Energy Commission fellowships in 1949. That was the earliest year in which selection instruments were used which closely resemble those currently employed in connection with the National Science Foundation fellowship program.

Among the selection techniques to be validated are Confidential Reports supplied by the candidates' professors, employers, etc. These Confidential Reports consist of a series of graphic rating scales in which the reporter checks his evaluation along a quantitative scale, and portions calling for descriptive comments on the candidates' qualifications. A graphic rating score can be derived by averaging the several graphic ratings, or by considering only the final summary or over-all scientific ability rating. Both methods have been used, and the results are usually so similar as to be interchangeable. In the case of the 1949 applications, it was felt most desirable to derive an average of all the ratings, because of the manner in which the form was set up, and the desire to have as wide a range of factors as possible represented in the graphic rating. Figure 1 presents in condensed form the pertinent portions of the 1949 Confidential Report form.

Figure 1
AEC PREDOCTORAL CONFIDENTIAL REPORT FORM
(condensed)

1. How well and under what association have you known the candidate?

2. Does the candidate clearly have the physical, personal, and intellectual qualities to earn the Ph.D. degree?

3. Where do you rate the candidate among all the graduate students in the same field, at his level of training and with whose work you have been associated during the past ten years? Indicate your candid evaluation by placing one check mark in each row of the table below. How many are in the group with which the candidate is being compared?_____.

	No Opportunity to observe	Below Average	Average	Among Best 25%	Among Best 10%	Among Best 5%	Best 2% Truly Exceptional
General all around ability							
Ability to read and evaluate scientific literature in his field							
Ability to organize and prosecute a research program effectively							
Ability to evaluate results critically and draw complete and correct conclusions							
Capacity for clear self expression							
Vigor and Initiative							
Dependability and self-reliance							

4. Give, if possible, specific evidence supporting the ratings you have given the candidate in Question 3.

5. How does this candidate compare with any present or previous National Research Council Fellows you have known? Mention fellows by name.

6. If more than one candidate is applying from your department, please give a comparative rating.

7. Is the candidate the kind of person with whom you would like to be closely associated as a professional colleague? If you check any square except Highly Acceptable, please include comment under Question 8 below.

Highly Acceptable ☐ Acceptable ☐ Not entirely Acceptable ☐ Would be Tolerated ☐

8. Give your judgment of the candidate's character and personality, including a statement of any physical, personality or character handicaps.

9. State any other facts you think important for the Fellowship Board to know when evaluating this candidate's qualifications, including concrete evidence of his creative ability in scientific work, appraisal of orderliness, rate of production, laboratory "know-how", persistence, versatility, regularity, etc.

Although the format of the Confidential Report has varied from year to year, the major characteristics have been similar throughout the past eight years. A difference worthy of note in connection with the current (i.e. 1955 and 1956) report is that Questions 2, 4, and 7 have been dropped, and more unstructured space made available for the generalized Question 9.

The graphic rating score is computed by averaging the scale values for all the rating scales in which scorable check marks are placed. Some raters omitted certain scales, some checked "no opportunity to observe". These scales were omitted in computing the average. The scale value was set at zero for the "below average" category, one for "average" etc. up to five for "truly exceptional". The range of possible graphic rating scores was thus zero through five.

In addition to the graphic rating score, mechanically derived as described from the check marks on the graphic scales, an additional score was derived from the free response or descriptive portions of the report form. The remainder of this report is concerned with the derivation of this descriptive report score, and evaluation of its characteristics and relationships to the graphic score. These evaluations were necessary in order to answer the following questions:

a. Can such a score be derived reliably and economically?
b. Do Confidential Reporters agree among themselves in the evaluations which are indicated by these scores?
c. Do these scores yield unique information--i.e., something not already indicated by the simpler graphic rating scales?

Quantitative answers were sought to each of these questions, in order to decide whether such a score would be worth carrying through the validation procedure, and to provide data necessary for adequate interpretation of the validity data when it is secured.

The standardization of the technique for deriving the Descriptive Report Score was the first step in determining its reliability. A set of scale steps was constructed against which any descriptive report could be evaluated. This relatively simple procedure turned out to be very satisfactory, in that it was found that most of the reports could be interpreted in terms of a quantitative scale of general scientific ability. The scale is as follows:

Description of Scale Step	Symbol	Numerical Value
Excellent student. No doubt of his success in attaining the PHD and in professional work thereafter	A	4.0
Very Good. A few weak spots, perhaps, but well compensated. Distinctly above the average PhD candidate	B	3.0
Average PhD candidate, all things considered	C	2.0
Below Average. Serious doubt of success. Might get PhD, but should be a terminal Master at best	D	1.0
Not suited for graduate school work	E	0.0

Next, rules for use of this quantitative scale were formulated: The evaluation was to be made according to the judged intent of the Confidential Reporter, not according to the estimate of success by the judge who scored the report. Intermediate points on the quantitative scale, e.g., 2.5 or B-, could be assigned if desired. In assigning these values, evidence would be sought of the Confidential Reporter's meaning in other than the graphic rating scales. The graphic scales per se were to be ignored. Examples of accomplishments were to be weighted heavily. Comparison with other NRC Fellows were crucial, when made. As these NRC Fellows were at the postdoctoral level and were carefully selected, a judgment that a candidate was equal to such Fellows would automatically rate him an "A". Each judge was to seek, in assigning quantitative values, to express the Confidential Reporter's intent, stripped of semantic variations between reporters. Such were the ground rules under which the evaluations were made.

The judges in this instance were selected clerical personnel of the NRC Fellowship office. All college graduates of high intelligence, they were specially trained in this particular evaluation system. One series of reports was used for initial practice, and the results thoroughly critiqued. A second series was then used for check evaluations, to insure that a uniform understanding of the evaluation process and ground rules had been reached. The extent to which this training was effective can be measured by the "scoring reliability" of the descriptive report scores.

Each judge evaluated each Confidential Report independently. The extent to which these judges arrived at corresponding evaluations of the reporter's intent may be termed the "scoring reliability" of the Descriptive Report Score for any individual report. (The term "scoring reliability" is used to distinguish between this degree of agreement between judges regarding a single report and that of different reporters regarding the same individual. The latter question will be dealt with systematically later.) The degree of agreement between each pair of judges (A vs. B, B vs. C and A vs. C) was computed. These three coefficients were found to be

similar in magnitude, and the statistics for all three inter-comparisons were combined to yield what is known as an intra-class correlation. It expresses the degree of agreement to be expected between any two such judges taken at random. This correlation coefficient, .82, does not indicate the reliability or stability of the average of three independent scorings, however. Because three such scorings were made, the reliability of the average score thus derived was estimated by application of the Spearman-Brown Prophecy Formula. The resulting coefficient, .93, is highly satisfactory. Further analysis could therefore be undertaken with confidence in the accuracy of scoring of the Confidential Reports.

The reliability of scoring thus established, the next question concerned economy. Although precise time records were not kept, it is known that from one to several minutes is required for the reading and evaluation of each report. If two minutes is assumed as a minimum figure, and three reports per applicant is the average, then with three separate scorings, 13 minutes per candidate is the estimated scoring time for these reports. It is apparent that this is a process which could be justified only on a small research scale, and that it would be completely infeasible in a large operating program. Even if large numbers of capable judges could be found and their employment financed, it is highly probable that quality control, or comparability of subjective scoring standards, could not be established and maintained. On the other hand, the possible merits of this approach for research purposes are such as to fully justify an initial study.

The cases selected for this pilot study comprised all the 1949 applicants for predoctoral fellowships in biological sciences for whom complete data were available. They numbered 215, including 99 awardees and 116 to whom no awards were made. A minor percentage of each group was composed of 1948 fellows applying for renewals.

The degree of agreement between Confidential Reporters was the next question. Typically, in large scale operations in which judgments of people are rendered by other people of miscellaneous background, the degree of agreement is not high. Even when raters are systematically trained together to minimize semantic difficulties, wide variation in judgments result if the objects of judgment are other people. One place where such judgments have been studied extensively is the U. S. Army. Judgments made for officer efficiency rating purposes are of vital importance, both to the Army as a whole and to the individuals being rated. Elaborate procedures are used to try to insure as uniform a basis for judgment as possible, in order that rating scores may reflect individual differences in the rated officers, rather than variations in standards applied by those rendering the ratings. Yet even here the degree of agreement is low. It was not surprising, therefore to find that the Confidential Reporters, of varying ages and degrees of experience, in all types of institutions of higher education throughout the

country, and with only the instructions on the form as a guide, should have a relatively low degree of agreement in evaluating these fellowship applicants. The intra-class correlation coefficient for the agreement of any two Confidential Reporters taken at random was found to be .42.

The reliability of the Descriptive Report Score for a single Confidential Reporter does not furnish all the pertinent information with regard to this score. In practice, there is almost always more than one reporter, and usually three. The very fact that they disagree among themselves, reflects in part the further fact that they perceive and evaluate different aspects of the candidate's abilities and personality traits. This is most obviously true of the free-response questions that form the basis for the Descriptive Report Score. One may hope that three reporters would, together, perceive most of the pertinent aspects of a candidate's qualifications, and that the sum of their evaluations would agree better with that of another comparable panel of three than would any pair of such reporters taken at random. The reliability of the summated judgment of such a panel of three (or of any other number) can be estimated by the Spearman-Brown Prophecy formula. As can be seen in Table 1, where the pertinent statistics are summarized, the reliability of the Descriptive Report Score average for three reporters is .68. This is a significant boost from the individual reliability figure of .42 and indicates a score which has quite satisfactory reliability for research purposes. To furnish a point of reference against which to judge this coefficient, the reliability of the graphic rating scores was similarly investigated.

Table 1

Means, Standard Deviations, and Intra-Class Reliability Coefficients of the Descriptive Report Scores

Group	Number of Cases	Number of Ratings	Mean Score	S.D. of Scores	Reliabilities* r_{11}	r_{33}
Awardees	99	324	3.4	.64	.21	.44
Rejectees	116	310	2.9	.90	.42	.68
Total Group	215	634	3.2	.89	.42	.68

*r_{11}=correlation between any two reporters taken at random

r_{33}=reliability of the average of three reporters, as estimated by the Spearman-Brown Prophecy Formula: $r_{nn} = \frac{n\, r_{11}}{1 + (n-1)\, r_{11}}$

where n = number of reporters per candidate

r_{nn} = reliability of the composite of n Confidential Rpts.

The graphic rating scores used in this study, as mentioned earlier, were derived by averaging the ratings rendered on a series of scales, omitting those checked "no opportunity to observe" and any scales not checked at all. Each reporter obviously made his own decisions as to which characteristics he had had an opportunity to observe. Those sets of scales which were incomplete, therefore, might theoretically be measuring different sets of variables than would be reflected in a completely filled out set of rating scales. A considerable body of experimental evidence exists, however, which indicates that these variations are unimportant in extent -- at least in comparison with the wide range of variation in rating standards which reporters attach to any set of scales. This evidence is largely derived from the fact that ratings on sets of traits are highly intercorrelated, even when strenuous efforts are made to eliminate overlapping, and when from the semantic point of view the scales appear unique. In any case the variation due to incomplete sets of scales was for the purposes of this experiment judged to be unimportant, and was disregarded.

It will be noted that the "scoring reliability" of the graphic scales is practically perfect, as the process is completely mechanical. Scoring time, also is low -- a matter of a few seconds per report. The agreement between reporters, however, is low (.31 for a single report, .57 for the average of three) as shown in Table 2.

Table 2

Means, Standard Deviations, and Intra-Class Reliability Coefficients of the Graphic Rating Scales (Total Score)

Group	Number of Cases	Number of Ratings	Mean Score	S.D. of Scores	Reliabilities* r_{11}	r_{33}
Awardees	99	338	3.48	.90	.28	.54
Rejectees	116	316	2.88	1.07	.21	.44
Total Group	215	654	3.19	1.03	.31	.57

*r_{11} and r_{33} have the same meanings as in Table 1

The fact that it is even lower than with the Descriptive Report Score may be largely an artifact, arising from a problem partly inherent in that score and partly a result of practical economic considerations. In a few instances, the descriptive comments were keyed to the graphic ratings in such a fashion as to be meaningless without them. The judge then must know the graphic score in order to derive the Descriptive Report Score. The reverse is never true with the scoring procedure here used. As a practical matter, all the data on each candidate -- graphic and descriptive -- were present in one folder used by the judges. Some contamination must therefore be expected, resulting in some spurious rise in the

reliability of the Descriptive Report Score. The training of the judges and the rules under which they operated certainly minimized this effect; it is judged not to be extensive, or else the reliability of the Descriptive Report Score would be far higher than the coefficient actually obtained.

The third and crucial question to which an answer was sought was whether the Descriptive Report Score yielded unique information about the candidate. It is obvious that the descriptive comments in their original form do contain information not indicated by the check marks on the graphic rating scales. That is a different question, however. Much of this added information is corroborative detail. When scored, the quantitative rating derived would be expected to agree rather well with the graphic rating which it was intended to corroborate. Thus no unique quantitative information would be provided. Another type of information found in the descriptive comments relates to the circumstances in which the Confidential Reporter formed his judgment. This information might add to (or in some cases subtract from) the confidence one might place in the quantitative score. But of itself it provides no scorable evidence regarding the candidate. A third type of information (fortunately rare) is that which is simply irrelevant to the point at issue. It might provide some information in an oblique way, with regard to the reporter. But it tells nothing about the candidate.

The fourth type of information which is provided by the descriptive comments is evidence regarding the candidate's ability or manner of functioning which is relevant to scientific ability but which is not necessarily reflected in the graphic rating scales. This might occur because none of the graphic scales relates to the particular type of ability described. It may be argued that the scale labeled "general all-around ability" should reflect this specific trait with its proper weighting in the whole complex of abilities. However, this is an interpretation which certainly would not be universally accepted. It can also be argued that while this interpretation might not be accepted when stated expressly, it will nevertheless in practice be acted upon, consciously or otherwise, by a Confidential Reporter filling out a rating scale. This, however, is but an assumption, and, in a nut-shell, it is this assumption which this research study seeks to test.

The extent to which the Descriptive Report Score yields information not contained in the graphic report score may be assessed by measuring the extent to which the correlation between the two scores approximates the correlation which would be obtained if the two scores were perfectly reliable. In fact, as we have seen, the two types of scores are attenuated by a considerable admixture of variable error. In order to compute a correction factor for this attenuation, independent determinations of the error and of the correlation between the two scores are necessary. Under the

conditions of this experiment complete independence of these determinations is difficult to obtain. A degree of independence is secured, however, by correlating the Descriptive Report Score of one reporter with the graphic rating rendered by another reporter, and by correlating the two scores rendered by the same reporter. From these data a correction for attenuation is derived. This correction is based on the principle that if two variables measure the same basic continuum, though in different manners, they would correlate perfectly if they were perfectly reliable. The correlation will fall below 1.00, however, in proportion as the variables involved are attenuated by random error. With due allowance for this error, the coefficient is corrected back to 1.00. If the correction does not bring the coefficient to 1.00, the assumption is that the two variables do not measure the same continuum.

In practice this correction for attenuation is not perfect, however. The correction itself is always subject to random error. In the present case, as mentioned above, it is also subject to a systematic error of unknown but probably small magnitude. When this correction for attenuation is applied, the corrected correlation coefficient is .75 for the total group under consideration. This coefficient is given in Table 3, together with the other data used in its derivation.

Table 3

Relationships Between Graphic Rating Score and Descriptive Report Score

Cases Involved in Correlation	Correlation Symbol *	Awardees	Rejectees	Total Group
Some reporters, different scores	$r_{g_1d_1}$	.50	.64	.63
Different reporters, different scores	$r_{g_1d_2}$	.19	.22	.27
Coefficient corrected for attenuation	r_{x^1gd}	.90	.65	.75

*g=graphic rating score
d=descriptive report score
1=any reporter
2=all reporters other than #1

Because the correction factor is large in this case, its own error is large. This introduces a difficulty in determining whether the obtained difference between the corrected coefficient of .75 is really significantly different from 1.00. The difficulty is compounded by the fact, previously mentioned, that the error estimate for the correction factor is not unbiased in this experiment. However, the best probability is that the obtained coefficient reflects a real underlying difference between these two types of scores.

Certainly the hypothesis that there is no difference cannot be confidently accepted. Thus, although a clean cut statistical answer cannot be obtained in this case, a practical operational one can be: The case for different fundamental variables is good enough to warrant carrying both types of scores to the validation stage.

In summary, the answer to the third question of our original set is that the probability exists that the Descriptive Report Score does measure something beyond what is measured by the graphic rating score. The extent of this probability is sufficient to warrant independent validation of both scores. The nature of the difference is not revealed by the present experiment. The correlation of each of these variables with various types of criterion data to be collected will throw some light on the nature of the two rating scores. These analyses are expected to be the subject of later reports.